Gaussian Process Regression for Voice Activity Detection and Speech Enhancement

Sunho Park, Seungjin Choi

***Abstract*— Gaussian process (GP) model is a flexible nonparametric Bayesian method that is widely used in regression and classification. In this paper we present a probabilistic method where we solve voice activity detection (VAD) and speech enhancement in a single framework of GP regression, modeling clean speech by a GP smoother. Optimized hyperparameters in GP models lead us to a novel VAD method since learned length-scale parameters in covariance functions are much different between voiced and unvoiced frames. Clean speech is estimated by posterior means in GP models. Numerical experiments confirm the validity of our method.**

I. Introduction

SPEECH enhancement is a fundamental processing, the goal of which is to estimate clean speech, given noise-contaminated signals. Signals measured through microphones in real-world environments, are always noisy data, hence, the enhancement of speech or the elimination of noise, plays a critical role for successful subsequent speech processing. Various methods have been developed for speech enhancement, including Wiener filter method [7], spectral subtraction method [1], HMM-based method [13], signal subspace method [3], Kalman filter method [9], H_∞ filter-based method [15], particle filter method [10], [16]. Speech enhancement methods often require voice activity detection (VAD) so that speech enhancement is applied to only voiced frames in order to save the computational load. In general, VAD and speech enhancement have been separately studied and two different methods are jointly applied in a cascade manner.

Gaussian process (GP) model has been widely used in machine learning because of its flexible nonparametric nature and computational simplicity [11], [14], [17]. In this paper we present a GP model to solve VAD and speech enhancement in a single framework. To this end, in each frame of length N ($N = 160$ in our experiments, corresponding to 20 ms duration in the case of 8 kHz sampling frequency), we model the clean speech by a latent function that takes past p and future p samples as inputs. Such a latent function is represented by random process with Gaussian prior. That is, we formulate speech enhancement as a GP regression problem. In the hyperparameter learning in GP models, Gaussian kernel widths (length-scale parameters in squared exponential covariance function) are much different between voiced and unvoiced frames. This enables us to detect voice activity frames. The clean speech is estimated by posterior means in GP models. Numerical experiments are provided, confirming the useful behavior of our GP method for VAD and speech enhancement.

Sunho Park and Seungjin Choi are with the Department of Computer Science, Pohang University of Science and Technology, Korea (phone: +82-54-279-2259; fax: +82-54-279-2299; email: {titan,seungjin}@postech.ac.kr).

II. Gaussian Process Model for Speech

We assume that the noisy speech signal x_t is a sum of clean speech s_t and the white Gaussian noise n_t (with mean 0 and variance σ_n^2):

$$x_t = s_t + n_t, \tag{1}$$

where $n_t \sim \mathcal{N}\left(0, \sigma_n^2\right)$. In each frame of length N, x_t is assumed to be a stationary process. We model the clean speech by a latent function involving past p samples and future p samples of x_t, leading to

$$x_t = f(\boldsymbol{x}_{\setminus t}) + n_t, \tag{2}$$

where

$$\boldsymbol{x}_{\setminus t} = [x_{t+p}, x_{t+p-1}, \ldots, x_{t+1}, x_{t-1}, x_{t-2}, \ldots, x_{t-p}]^\top.$$

The model (2) is independently applied to each frame, i.e., the nonlinear functions $f(\cdot)$ are different across frames. With abuse of notations, we use $f(\cdot)$ without specifying the frame index.

GP model represents the latent function $f(\cdot)$ by a random process with Gaussian prior, i.e.,

$$f(\boldsymbol{x}_{\setminus t}) \sim \mathcal{GP}(0, k(\boldsymbol{x}_{\setminus t}, \boldsymbol{x}_{\setminus \tau})), \tag{3}$$

$k(\cdot, \cdot)$ is a covariance function. We use the squared exponential covariance function (Gaussian kernel),

$$k\left(\boldsymbol{x}_{\setminus t}, \boldsymbol{x}_{\setminus \tau}\right) = \exp\left\{-\|\boldsymbol{x}_{\setminus t} - \boldsymbol{x}_{\setminus \tau}\|^2 / l\right\}, \tag{4}$$

where $\|\cdot\|$ denotes Euclidean norm and $l > 0$ is a length-scale parameter (kernel width parameter).

Given a collection of input $\boldsymbol{X} = \left\{\boldsymbol{x}_{\setminus t}\right\}_{t=\xi+1}^{\xi+N}$ ($\xi = N(m-1)$ for the mth frame) and hyperparameters $\boldsymbol{\theta} \triangleq [l, \sigma_n^2]^\top$, the prior of latent functions is given by

$$p(\boldsymbol{f} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{N}(0, \boldsymbol{K}), \tag{5}$$

where

$$\begin{aligned} \boldsymbol{f} &= \left[f(\boldsymbol{x}_{\setminus(\xi+1)}), ..., f(\boldsymbol{x}_{\setminus(\xi+N)})\right]^\top, \\ [\boldsymbol{K}]_{u,v} &= \exp\left\{-\|\boldsymbol{x}_{\setminus(\xi+u)} - \boldsymbol{x}_{\setminus(\xi+v)}\|^2 / l\right\}. \end{aligned}$$

Define response variables by

$$\boldsymbol{y} = [x_{\xi+1}, x_{\xi+2}, \ldots, x_{\xi+N}]^\top.$$

Then, the likelihood of $\boldsymbol{y}$ given the latent functions $\boldsymbol{f}$ is derived from the observation model (1),

$$p(\boldsymbol{y} \mid \boldsymbol{f}, \boldsymbol{\theta}) = \mathcal{N}(0, \sigma_n^2 \boldsymbol{I}), \tag{6}$$

where $\boldsymbol{I}$ is the $N \times N$ identity matrix. The required tasks for speech enhancement are two folds: (1) the estimation of optimal hyperparameters $\widehat{\boldsymbol{\theta}}$ (leading to VAD); (2) the calculation of the posterior mean of $\boldsymbol{f}$ (leading to clean speech estimation).

III. The Proposed Method

A. Voice Activity Detection by Length-Scale Parameter

Our GP model-based VAD method (GP-VAD) considers optimized length-scale parameters (kernel width) to detect voiced frames. The GP-VAD method can be viewed a special case of automatic relevance determination (ARD) that automatically selects the relevant dimensions of input variables [8]. In ARD, the relevance of each dimension is determined by the inverse of length-scale parameters $\{1/l_1, \ldots, 1/l_{2p}\}$. When the ith length-scale is very large, the covariance function is almost independent of the ith input and its contribution to inference is automatically removed. We directly apply this simple idea to the task of VAD, observing that learned length-scale parameters are very large for unvoiced frames (see Fig. 1).

In an unvoiced frame, GP smoother produces $f(\boldsymbol{x}_{\xi+i}) \approx 0$ for $i = 1, \ldots, N$, leading to $k(\boldsymbol{x}_{\setminus t}, \boldsymbol{x}_{\setminus \tau}) \approx 1$. Thus the length-scale parameter becomes considerably large in such a case. Unvoiced frames are easily detected by monitoring learned length-scale parameters, i.e., when the learned length scale parameter is larger than the threshold l_{thr}, it is decided as an unvoiced frame. Fig. 1 shows an illustrative example. In the 3rd plot in Fig. 1, length-scale parameters across frames are plotted, where parameters are considerably large for unvoiced frames. On the other hand, in the entropy-based method (En-VAD) [12], entropies are used (see the 4th plot in Fig. 1), where a careful selection of threshold is required [2].

In our GP model, hyperparameters for each frame are learned by maximizing the marginal likelihood that is of the form [11]:

$$\begin{aligned} \log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}) &= \int p(\boldsymbol{y} \mid \boldsymbol{f}, \boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{f} \mid \boldsymbol{X}, \boldsymbol{\theta}) d\boldsymbol{f} \quad (7) \\ &= -\frac{N}{2} \log 2\pi - \frac{1}{2} \boldsymbol{y}^\top (\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} \boldsymbol{y} \\ &\quad -\frac{1}{2} \log |(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})|, \quad (8) \end{aligned}$$

where $|\cdot|$ is the determinant of a matrix. A gradient-based method can be used to estimate hyperparameters $\boldsymbol{\theta} = [l, \sigma_n^2]^\top$. Note that parameters l and σ_n^2 parameters should be positive. Thus the optimization (8) with respect to $\boldsymbol{\theta}$ is actually a constrained optimization. In practice, this optimization is easily solved by an unconstrained optimization with respect to the logarithm of hyperparameters, $\{\log l, \log \sigma_n^2\}$. The gradient of (8) with respect to the ith element of $\log \boldsymbol{\theta}$ is given by

$$\begin{aligned} &\frac{\partial}{\partial \log \theta_i} \log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}) \\ &= \frac{1}{2} \boldsymbol{y}^\top \boldsymbol{M}^{-1} \frac{\partial \boldsymbol{M}}{\partial \log \theta_i} \boldsymbol{M}^{-1} \boldsymbol{y} - \frac{1}{2} \mathrm{tr}(\boldsymbol{M}^{-1} \frac{\partial \boldsymbol{M}}{\partial \log \theta_i}), \quad (9) \end{aligned}$$

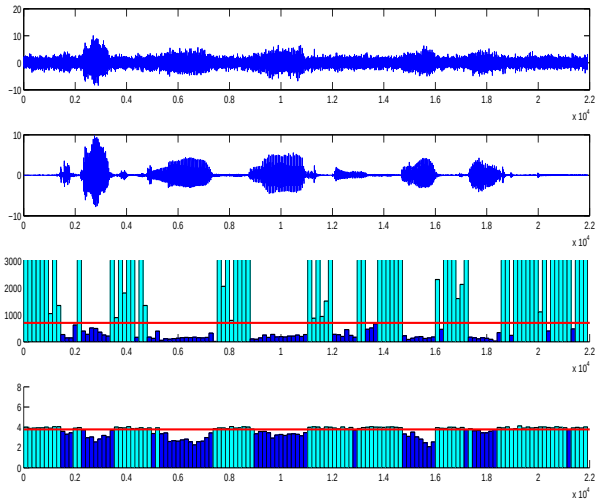

Fig. 1. From top to bottom: a noise-contaminated speech signal (at input SNR 0dB); original clean speech; voiced/unvoiced frames detected by our GP-VAD method; voiced/unvoiced frames detected by the entropy-based method (En-VAD).

where θ_i is the ith element in $\boldsymbol{\theta}$ and $\boldsymbol{M} = \boldsymbol{K} + \sigma_n^2 \boldsymbol{I}$. For VAD, we need $\partial \boldsymbol{M} / \partial \log l$ that is given by

$$\frac{\partial \boldsymbol{M}}{\partial \log l} = (1/l) \boldsymbol{G} \odot \boldsymbol{K}, \quad (10)$$

where $\boldsymbol{G}$ is a squared Euclidean distance matrix whose (u, v)-element is given by

$$[\boldsymbol{G}]_{u,v} = \|\boldsymbol{x}_{\setminus(\xi+u)} - \boldsymbol{x}_{\setminus(\xi+v)}\|^2,$$

and $\odot$ is Hadamard product(elementwise multiplication). In the case of noise variance, the derivative is given by

$$\frac{\partial \boldsymbol{M}}{\partial \log \sigma_n^2} = \sigma_n^2 \boldsymbol{I}. \quad (11)$$

With these gradient calculations, any gradient-based optimization methods can be applied to learn hyperparameters that optimize the marginal likelihood. We use *fminunc* in Matlab optimization toolbox in this paper.

B. Estimation of Clean Speech

In a voiced frame, we need to estimate clean speech from noisy observations. In our GP model, the clean speech is estimated from posterior distribution of the latent functions $\boldsymbol{f}$, which is calculated by

$$p(\boldsymbol{f} \mid \boldsymbol{y}, \boldsymbol{X}, \widehat{\boldsymbol{\theta}}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{f}, \boldsymbol{X}, \widehat{\boldsymbol{\theta}}) p(\boldsymbol{f} \mid \boldsymbol{X}, \widehat{\boldsymbol{\theta}})}{p(\boldsymbol{y} \mid \boldsymbol{X}, \widehat{\boldsymbol{\theta}})}. \quad (12)$$

It follows from the Gaussian likelihood (6) that (12) is easily computed by

$$p(\boldsymbol{f} \mid \boldsymbol{y}, \boldsymbol{X}, \widehat{\boldsymbol{\theta}}) = \mathcal{N}(\overline{\boldsymbol{f}}, \overline{\boldsymbol{\Sigma}}), \quad (13)$$

where the posterior mean $\overline{\boldsymbol{f}}$ and the posterior covariance matrix $\overline{\boldsymbol{\Sigma}}$ are calculated by,

$$\overline{\boldsymbol{f}} = \boldsymbol{K}(\boldsymbol{K} + \widehat{\sigma}_n^2 \boldsymbol{I})^{-1}\boldsymbol{y}, \quad (14)$$

$$\overline{\boldsymbol{\Sigma}} = \boldsymbol{K} - \boldsymbol{K}(\boldsymbol{K} + \widehat{\sigma}_n^2 \boldsymbol{I})^{-1}\boldsymbol{K}. \quad (15)$$

The estimation of the clean speech only needs the posterior mean $\overline{\boldsymbol{f}}$. The estimator of clean speech $\hat{s}_t$ is

$$\hat{s}_{\xi+i} = \overline{\boldsymbol{f}}_i, \quad (16)$$

where $\overline{\boldsymbol{f}}_i$ means the ith element of $\overline{\boldsymbol{f}}$.

The result (16) is understood as the smoothing of the response variables $\boldsymbol{y}$. Let $\{\lambda_i, \boldsymbol{u}_i\}_{i=1}^N$ be the eigenvalues and eigenvectors of $\boldsymbol{K}$. The response $\boldsymbol{y}$ can be represented by the set of the eigenvectors, $\boldsymbol{y} = \sum_{i=1}^N \gamma_i \boldsymbol{u}_i$, where $\gamma_i = \boldsymbol{u}_i^\top \boldsymbol{y}$. Then (16) is rewritten by

$$\overline{\boldsymbol{f}} = \sum_{i=1}^{N} \frac{\gamma_i \lambda_i}{\lambda_i + \widehat{\sigma}_n^2} \boldsymbol{u}_i. \quad (17)$$

We easily check that the component in $\boldsymbol{y}$ along $\boldsymbol{u}_i$ is eliminated if $\lambda_i/(\lambda_i + \widehat{\sigma}_n^2) \ll 1$ [11]. In the cases of most covariance function, the larger eigenvalues correspond to more slowly varying eigenvectors. Thus high-frequency components in $\boldsymbol{y}$ is removed [11]. It is analog that high frequencies of noisy speech are smoothed out in frequency domain.

TABLE I

ALGORITHM OUTLINE: OUR GP MODEL-BASED SPEECH ENHANCEMENT.

Let N be the length of each frame, m be the index of frames and $\xi = N * (m-1)$.
for $m = 1, ..., M$
  Set $\boldsymbol{y} = \left[x_{\xi+1}, x_{\xi+2}, \ldots, x_{\xi+N}\right]^\top$,
  $\boldsymbol{X} = \{\boldsymbol{x}_{\setminus t}\}_{t=\xi+1}^{\xi+N}$,
  Find $\hat{l}$ and $\hat{\sigma}_n^2$ (see Sec. III-A),
  if $\hat{l}$ is grater than $\hat{l}_{thr}$:
    Set $\{\hat{s}_t\}_{t=\xi+1}^{\xi+N}$ to random noise close to zero,
  otherwise:
    Construct the kernel matrix $\boldsymbol{K}$ with $\hat{l}$ (5),
    Estimate $\{\hat{s}_t\}_{t=\xi+1}^{\xi+N}$ using (14) and (16),
end

IV. NUMERICAL EXPERIMENTS

We perform experiments using 15 different speech samples taken from NOIZEUS corpus[1], each of which is about 2-3 seconds long and is resampled at 8 kHz. White Gaussian noise is synthetically added to clean speech in order to generate noise-contaminated signals under various input SNR conditions ranging from -2 dB to 8 dB. In all experiments we use:

- $p = 10$
- $\hat{l}_{thr} = 0.7 \times 10^3$ (the threshold involving the length-scale parameter for VAD)
- $N = 160$ (the length of each frame, associated with 20 ms long)

Experiments are performed for two tasks: (1) VAD; (2) speech enhancement. In the case of VAD, we compare our GP method (GP-VAD) to the entropy-based method (En-VAD) [12], in terms of receiver operating characteristics (ROC) cure and area under curve (AUC) [4]. Fig. 2 show ROC curves at different SNR levels for our method and En-VAD. Associated AUC is summarized in Table II. At SNR levels considered in experiments, our GP method outperforms En-VAD, since the length-scale parameters in our GP method are considerably different between voiced and unvoiced frames.

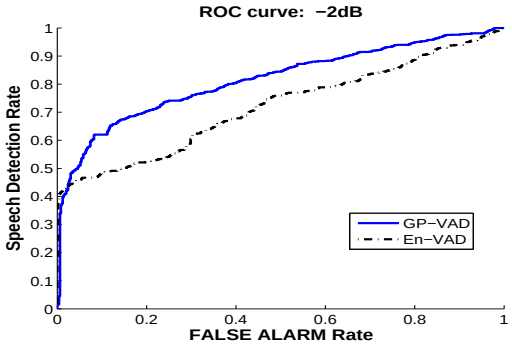

Fig. 2. ROC curves for our method (GP-VAD) and entropy-based method (En-VAD) in the case of input SNR = -2 dB.

TABLE II

AUC FOR OUR METHOD (GP-VAD) AND ENTROPY-BASED METHOD (EN-VAD) UNDER VARIOUS NOISE CONDITIONS.

Input SNR	GP-VAD	En-VAD
-2 db	**0.8151**	0.7190
0 db	**0.8262**	0.7844
2 db	**0.8484**	0.7994
4 db	**0.8835**	0.8327
6 db	**0.8752**	0.8472
8 db	**0.9033**	0.8745

Next we evaluate the speech enhancement performance of our GP-based method that is presented in Sec. III-B, with comparison to Kalman filter method [5]. After VAD is done, we apply our GP method to voiced frames, in order to estimate clean speech through posterior means of latent functions. We consider two different measures to evaluate the performance: (1) root mean squared log-spectral-distance (rmsLSD) [6]; (2) output SNR. The rmsLSD reflects the spectral distance between true clean speech and estimated speech computed, which is defined by

$$\text{rmsLSD}^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |V(\omega)\ |^2 \, d\omega, \quad (18)$$

where $V(\omega)$ is the difference between two spectral models (clean speech and estimated speech), given by

$$V(\omega) = \ln(\sigma^2 / |\ A(e^{j\omega})\ |^2) - \ln((\sigma')^2 / |\ A'(e^{j\omega})\ |^2),$$

[1]Available at: http://www.utdallas.edu/.loizou/speech/noizeus/

where $A(e^{j\omega})$ and $A'(e^{j\omega})$ involve AR models of clean speech and estimated speech, respectively and σ and σ' represent the standard deviation of innovation (residual) in each model. The output SNR reflects the difference between clean speech and estimated speech in the time domain, which is defined by

$$\mathrm{SNR}_o \quad = \quad 10\log_{10}\frac{\sum_{t=1}^{T} s_t^2}{\sum_{t=1}^{T}[s_t-\widehat{s}_t]^2}, \qquad (19)$$

where $\widehat{s}_t$ is the estimated speech in (16).

Table III summarizes the performance of three methods: (1) our method; (2) Kalman filter (KF); Kalman filter with our GP-VAD (KF+VAD), showing that how much the output SNR and rmsLSD are improved, compared to the input SNR and rmsLSD (the higher SNR and the lower rmsLSD indicate better performance). In each performance measure, our GP method outperforms KF and KF+VAD, producing higher output SNR and lower rmsLSD.

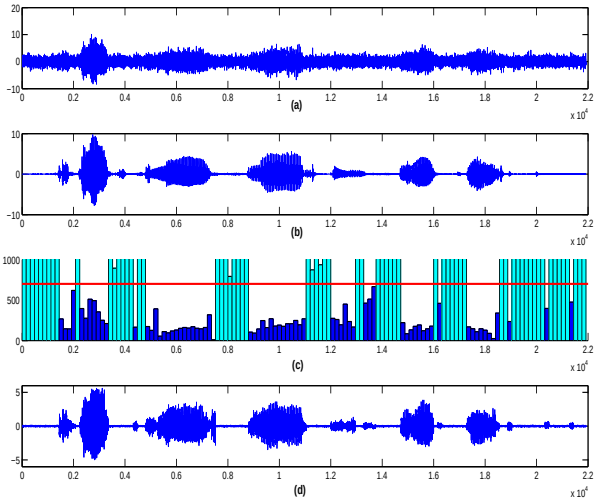

Fig. 3. The result of VAD and speech enhancement of our GP method is shown in the case of input SNR=0dB. From top to bottom: noisy speech signal; original clean speech; the result of VAD; enhanced speech by our GP method.

TABLE III

PERFORMANCE COMPARISON OF OUR GP METHOD TO KALMAN FILTER (KF) AND KALMAN FILTER WITH VAD (KF+VAD), IN TERMS OF SNR AND RMSLSD MEASURES. THE FIRST COLUMN REPRESENTS SNR AND RMSLSD BEFORE SPEECH ENHANCEMENT IS APPLIED.

(SNR, rmsLSD)	our method	KF	KF + VAD
(-2, 3.10) db	(**5.63**, **2.01**)	(3.06, 2.73)	(5.06, 2.04)
(0, 3.01) db	(**6.94**, **1.78**)	(4.38, 2.64)	(6.26, 1.94)
(2, 2.91) db	(**8.52**, **1.63**)	(6.06, 2.53)	(7.70, 1.86)
(4, 2.79) db	(**9.82**, **1.49**)	(7.74, 2.41)	(9.19, 1.73)
(6, 2.64) db	(**11.57**, **1.41**)	(9.52, 2.27)	(10.66, 1.71)
(8, 2.50) db	(**12.98**, **1.32**)	(11.33, 2.12)	(12.39, 1.57)

V. CONCLUSIONS

We have presented a method with GP models which jointly performed VAD and speech enhancement. Modeling clean speech by a GP smoother, we reformulated speech enhancement as a GP regression problem. Optimized length-scale parameters provided a solution to VAD since they are considerably different between voiced and unvoiced frames. Clean speech estimation was done by posterior means of latent functions in our GP models. Numerical experiments have shown that our GP method works better, compared to some of existing methods in the task of VAD and speech enhancement. Our current method works only when the noise is white Gaussian. Further extension is a future work that is being in consideration.

Acknowledgments: This work was supported by Korea Ministry of Knowledge Economy under the ITRC support program supervised by the IITA (IITA-2008-C1090-0801-0045).

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, 1979.

[2] D. V. Compernolle, "Noise adaptation in a hiddenMarkov model speech recognition system," *Computer Speech and Language*, pp. 151–167, 1989.

[3] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 355–358, Apr. 1993.

[4] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," HP Laboratories, Tech. Rep. HPL-2003-4, 2004.

[5] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 373–385, 1998.

[6] A. H. Gray Jr. and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 380–391, 1976.

[7] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 26, pp. 197–210, 1978.

[8] R. M. Neal, *Bayesian Learning for Neural Networks*. Springer, 1996.

[9] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1987, pp. 177–180.

[10] S. Park and S. Choi, "Rao-Blackwellized particle filtering for sequential speech enhancement," in *Proceedings of the International Joint Conference on Neural Networks*, Vancouver, Canada, 2006.

[11] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[12] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in *Proceedings of EUROSPEECH*, 2001.

[13] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 5, pp. 445–455, 1998.

[14] M. Seeger, "Gaussian processes for machine learning," *International Journal of Neural Systems*, vol. 14, no. 2, pp. 69–106, 2004.

[15] X. Shen and L. Deng, "A dynamic system approach to speech enhancement using the H_∞ filtering algorithm," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 4, pp. 391–399, 1999.

[16] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, pp. 173–185, 2002.

[17] C. K. I. Williams and C. E. Rasmussen, "Gaussian processes for regression," in *Advances in Neural Information Processing Systems*, vol. 8. MIT Press, 1996.